Assignment 3 - CSC/DSC 462 - Fall 2018 - Due November 6

Question 1 is worth 10 points, Questions 2 and 3 are worth 20 points each, for a total of 50 points.

Q1: The odds of an event A is denoted $Odds(A)$. Suppose the distribution of a random variable $X \in \{1, 2, 3, 4, 5\}$ depends on whether or not event A occurs. In particular, conditional on A, the PMF of X is given by $(p_1, p_2, p_3, p_4, p_5) = (0, 1/4, 1/4, 1/4, 1/4)$. Conditional on A^c, the PMF of X is given by $(p'_1, p'_2, p'_3, p'_4, p'_5) = (4/10, 3/10, 2/10, 1/10, 0)$.

Determine the relationship between $Odds(A \mid X = x)$ and $Odds(A)$ for $x = 1, 2, 3, 4, 5$. For which values of x does evidence of the form $\{X = x\}$ increase the odds that A occurs?

Q2: A test for a certain infection was evaluated experimentally. When administered to a test group of 425 individuals known to have the infection, the test was positive in 401 cases. The test was also administered to a control group of 765 subjects known to be free of the infection. The test was positive in 12 cases.

(a) Estimate the sensitivity and specificity of the test directly from the data.

(b) This test is intended to be used in clinical populations of varying infection prevalence. Use `R` to construct plots of PPV and NPV for values of prevalence ranging from 0 to 20%. Use the `type = 'l'` option of the `plot()` function.

(c) Calculate prevalence, NPV and PPV directly from the data. How do these values compare to those shown in the plots of part (b)?

(d) Give the relationship between the prior and posterior odds of infection for both a positive and negative test result.

Q3: For this question, we will make use of the notation of Section 5.4 of the lecture notes. In particular, a model for a diagnostic test for a certain condition relies on the four events:

$$
\begin{aligned}
O_- &= \{ \text{ the patient does not have the condition } \} \\
O_+ &= \{ \text{ the patient has the condition } \} \\
T_- &= \{ \text{ the patient tests negative } \} \\
T_+ &= \{ \text{ the patient tests positive } \}.
\end{aligned}
$$

The quantities *sens* (sensitivity), *spec* (specificity) and *prev* (prevalence) are defined as

$$
\begin{aligned}
sens &= P(T_+ \mid O_+) \\
spec &= P(T_- \mid O_-) \\
prev &= P(O_+).
\end{aligned}
$$

Suppose an existing diagnostic test is evaluated by estimating the conditional probabilities $P(O_+ \mid T_+)$ and $P(O_+ \mid T_-)$. For example, these conditional probabilities might be estimated by following up with additional testing patients who have tested positive and negative in a clinical setting.

We would like to verify that $P(O_+ \mid T_+) > P(O_+ \mid T_-)$, and then quantify this difference using some distance function. We will examine three such distances, in each case expressing the difference analytically as a function of *sens*, *spec* and *prev* (see Equations (5.6)-(5.7) in the lecture notes for reference).

(a) First consider the additive difference of the conditional probabilities:

$$\Delta = P(O_+ \mid T_+) - P(O_+ \mid T_-).$$

Express Δ as a function of *sens*, *spec* and *prev*. Does Δ depend on *prev*? In particular, what is the limit of Δ as *prev* approaches 0, and as *prev* approaches 1?

(b) Next, consider the *relative risk*, which is defined as the ratio of the conditional probabilities:

$$RR = \frac{P(O_+ \mid T_+)}{P(O_+ \mid T_-)}.$$

Express RR as a function of $sens$, $spec$ and $prev$. Does RR depend on $prev$? In particular, what is the limit of RR as $prev$ approaches 0, and as $prev$ approaches 1?

(c) The *odds ratio* is the ratio of the conditional odds:

$$OR = \frac{Odds(O_+ \mid T_+)}{Odds(O_+ \mid T_-)},$$

where, in general, the conditional odds is given by

$$Odds(A \mid B) = \frac{P(A \mid B)}{1 - P(A \mid B)}.$$

Express OR as a function of $sens$, $spec$ and $prev$. Verify that OR does not depend on $prev$.